Committee on Data Science

The Committee on Data Science (CDS) was established in 2023 to support graduate and undergraduate programs in this emerging discipline at the University of Chicago. Affiliated faculty come from numerous departments across campus with a core group in the departments of Statistics and Computer Science. CDS holds the educational philosophy that a strong program in Data Science should encompass foundational theory, methodological innovations and real-world applications. A Data Science education should draw from the intellectual tradition and key concepts of Computer Science, Applied Mathematics, Statistics, and other fields while providing a new integrative framework for data-driven thinking, discovery, and decision-making.

Committee on Data Science website: https://codas.uchicago.edu/

Committee Co-Directors

- Dan L Nicolae (Statistics)
- Michael J Franklin (Computer Science)

Program Faculty

- Luc Anselin (Sociology)
- Luis Bettencourt (Ecology, Sociology)
- Raul Castro Fernandez (Computer Science)
- Aloni Cohen (Computer Science)
- James Evans (Sociology)
- Nick Feamster (Computer Science)
- Robert Grossman (Medicine, Computer Science)
- Ari Holtzman (Computer Science, Data Science)
- Nikos Ignatiadis (Statistics, Data Science)
- Alex Kale (Computer Science, Data Science)
- Frederick Koehler (Statistics, Data Science)
- Sanjay Krishnan (Computer Science)
- Mina Lee (Computer Science, Data Science)
- Bo Li (Computer Science, Data Science)
- Tian Li (Computer Science, Data Science)
- Sendhil Mullainathan (Computation, Behavioral Science)
- Samantha Riesenfeld (Molecular Engineering, Medicine)
- Veronika Rockova (Econometrics, Statistics)
- Aaron Schein (Statistics, Data Science)
- Matthew Stephens (Statistics)
- Chenhao Tan (Computer Science, Data Science)
- David Uminsky (Computer Science)
- Blase Ur (Computer Science)
- Victor Veitch (Statistics, Data Science)
- Jingshu Wang (Statistics)
- Molly Offer-Westort (Political Science)
- Rebecca Willett (Statistics, CAMI, Computer Science)
- Haifeng Xu (Computer Science, Data Science)
- Ce Zhang (Computer Science, Data Science)

PhD in Data Science

Program Overview

The PhD in Data Science was developed to train all students in the mathematical foundations of data science, responsible data use and communication, as well as advanced computational methods. Candidates will be able to explore diverse research opportunities alongside distinguished Data Science faculty at UChicago.

Curriculum

The program requires students to complete nine courses: four required courses (1-4 below); one elective either in mathematical foundations or scalability and computing (5 or 6 below), and four other graduate-level electives that can come from proposed courses in Data Science or existing graduate courses in Computer Science

or Statistics. Some students, after consulting with the committee graduate advisor, might decide to take all nine courses over the first two years.

Required Courses:

1. Foundations in Machine Learning & AI - Part I
2. Responsible Use of Data & Algorithms
3. Data Interaction
4. Systems for Data and Computers / Data Design

Required Electives (Choose one of the following):

1. Foundations in Machine Learning & AI - Part II
2. Data Engineering & Scalable Computing

Thesis Advisor and Dissertation Committee

Students typically select a thesis advisor by the beginning of their second year. By the end of the third year, each PhD student shall establish a thesis committee of at least three faculty members, including the advisor, with at least half of the members coming from the Committee on Data Science (CDS).

Proposal Presentation & Admission to Candidacy

By the end of the third year, students should have scheduled and completed a proposal presentation to their committee in order to be advanced to candidacy. The proposal presentation is typically an hour-long meeting that begins with a 30-minute presentation by the student followed by a question and discussion period with the committee.

Admissions

The PhD in Data Science admits students each year for the Fall quarter only; a full list of admission requirements and a link to start your application can be found here (https://codas.uchicago.edu/how-to-apply/). If you have any questions regarding your application or the admissions process, please send your inquiry to data-science@uchicago.edu for a timely response.

Master's in Data Science (MSDS)

Program Overview

The Master's in Data Science (MSDS) was developed for students interested in pursuing a research career in Data Science with courses taught by faculty in Statistics, Computer Science, and other departments across the university.

Curriculum: Foundational Courses

The program offers three foundational courses. Students have the option to either (1) enroll in foundational courses in the summer before the program starts or (2) pass examinations to demonstrate proficiency in the material in lieu of enrolling in foundational courses.

The foundational courses are as follows:

1. Computational Foundations for Data Science

2. Mathematical Foundations for Data Science

3. Statistical Foundations for Data Science

Curriculum: Core & Elective Courses

In addition to the foundational courses (or passing examinations in lieu of enrollment in foundational courses), students must complete five required core courses, four graduate-level electives (approved by the Committee on Data Science), as well as a final project in order to be eligible for degree completion.

The core courses are as follows:

1. Introduction to Data Science
2. Systems for Data and Computers/Data Design
3. Data Interaction
4. Introduction to ML and AI or Foundations of Machine Learning and AI - Part I
5. Responsible Use of Data and Algorithms

Admissions

The Master's in Data Science (MSDS) admits students each year for the Fall quarter only; a full list of admission requirements and a link to start your application can be found here (https://codas.uchicago.edu/how-to-apply/). If you have any questions regarding your application or the admissions process, please send your inquiry to data-science@uchicago.edu for a timely response.

Data Science Courses

DATA 30100. Introduction to Data Science. 100 Units.
The course will focus on the analysis of real life data and on statistical and machine learning methods to perform inference and to predict future outcomes. It will cover topics from the whole data life cycle, ranging from data collection (including wrangling, cleaning, and sampling) to summarizing results through visualization and interpretable summaries, with a focus on extracting meaning, value and information from data. Important aspects in data science, such as bias, fairness, privacy while building algorithms and predictive models, will also be explored.
Instructor(s): D. Nicolae Terms Offered: Autumn
Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 30332. Thinking with Deep Learning for Complex Social & Cultural Data Analysis. 100 Units.
A deluge of digital content is generated daily by web-based platforms and sensors that capture digital traces of human communication and connection, and complex states of society, culture, economy, and the world. Emerging deep learning methods enable the integration of these complex data into unified social and cultural "spaces" that enable new answers to classic social and cultural questions, and also pose novel questions. From the perspective of deep learning, everything can be viewed as data-novels, field notes, photographs, lists of transactions, networks of interaction, theories, epistemic styles-and our treatment examines how to configure deep learning architectures and multi-modal data pipelines to improve the capacity of representations, the accuracy of complex predictions, and the relevance of insights to substantial social and cultural questions. This class is for anyone wishing to analyse textual, network, image or arbitrary structured and unstructured data, especially in concert with one another to solve complex social and cultural analysis problems (e.g., characterize a culture; predict next year's ideology).
Instructor(s): James Evans Terms Offered: Spring Winter
Prerequisite(s): The course uses Python and the widely popular PyData ecosystem to demonstrate all motivating examples and includes working code, accompanying exercises, relevant datasets and additional analytics and visualization that facilitate social and cultural interpretation and communication. Familiarity with Python is required.
Equivalent Course(s): MACS 27000, SOCI 30332, MACS 37000

DATA 31500. Data Interaction. 100 Units.
This course provides core knowledge and technical skills around data interfaces, with an emphasis on visualization and front-end software development. Graduate students in Data Science and Computer Science will engage in project-based learning to become fluent with visualization APIs, computational notebooks, web development, technical writing, and presentation. Topics of interest include data visualization design, spatial and visual reasoning, cartography, interactive articles, data storytelling, data-driven persuasion, uncertainty communication, and model interpretability.
Instructor(s): A. Kale Terms Offered: Autumn
Prerequisite(s): Consent of Instructor unless graduate student in Data Science
Equivalent Course(s): CMSC 31500

DATA 33221. Advanced Topics in Law and Computing. 100 Units.
This interdisciplinary seminar will bring together instructors and graduate students from Computer Science / Data Sciences and the Law School. The seminar's focus will be on topics where law and policy intersect with computer science. Such topics may include cryptography and encryption; electronic surveillance and criminal procedure; the Computer Fraud & Abuse Act; the law governing data breaches; redistricting and the US Census; deep fakes; GDRP, Europe's Digital Services Act and the CCPA; and international data transfers. Students will be evaluated on the basis of short bi-weekly reaction papers, class participation based on weekly assigned reading, and team projects that pair law students with computer and data scientists.
Equivalent Course(s): CMSC 33221

DATA 34100. Introduction to Data Systems and Data Design. 100 Units.
The goal of this course is to teach students: (1) how to think about data , its logical semantics, and what is a query; (2) how to practically handle data, both in relational databases and other more flexible data processing frameworks (e.g. Spark); (3) practical design principles about schema, integrity constraints, etc. (4) an introduction to systems that allows students to understand performance, and helps them become better users.
Instructor(s): C. Zhang Terms Offered: Autumn
Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 34200. Data Engineering and Scalable Computing. 100 Units.
This course covers the principles and practices of managing and processing data at scale. Students will learn about distributed systems, cloud computing, and big data technologies. Topics include data storage architectures, data catalogs and governance, distributed computing frameworks like Apache Spark, streaming data processing, and data transformation pipelines. The course will provide hands-on experience with state-of-the-art tools and

techniques for building end-to-end data engineering solutions to support large-scale data science, analytics and AI applications.
Instructor(s): M. Franklin Terms Offered: Winter
Prerequisite(s): DATA 34100; Consent of Instructor unless graduate student in Data Science

DATA 35422. Machine Learning for Computer Systems. 100 Units.
This course will cover topics at the intersection of machine learning and systems, with a focus on applications of machine learning to computer systems. Topics covered will include applications of machine learning models to security, performance analysis, and prediction problems in systems; data preparation, feature selection, and feature extraction; design, development, and evaluation of machine learning models and pipelines; fairness, interpretability, and explainability of machine learning models; and testing and debugging of machine learning models. The topic of machine learning for computer systems is broad. Given the expertise of the instructor, many of the examples this term will focus on applications to computer networking. Yet, many of these principles apply broadly, across computer systems. You can and should think of this course as a practical hands-on introduction to machine learning models and concepts that will allow you to apply these models in practice. We'll focus on examples from networking, but you will walk away from the course with a good understanding of how to apply machine learning models to real-world datasets, how to use machine learning to help computer systems operate better, and the practical challenges with deploying machine learning models in practice."
Instructor(s): Nick Feamster
Prerequisite(s): CMSC 14300 or CMSC 15400
Equivalent Course(s): DATA 25422, CMSC 35422, CMSC 25422

DATA 37000. Introduction to Machine Learning and Neural Networks. 100 Units.
This course is an introduction to machine learning (ML) for students to build a solid foundation in modeling and data science. It will cover both unsupervised and supervised ML algorithms, with the latter focusing on both regression and classification models. Python is the programming language of choice for implementing various models to solve complex problems across multiple domains. The course will also introduce basic neural network architectures, including Single-Layer Perceptron (SLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Students will apply these techniques in contexts where they are most effective. A strong understanding of linear algebra, multivariable calculus, and statistics/probability theory is expected. Python coding assignments and projects will be integral to the course.
Instructor(s): E. Lo Terms Offered: Autumn
Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 37100. Introduction to AI: Deep Learning and GAI. 100 Units.
Artificial Intelligence is transforming industries and daily life, permeating almost every aspect of modern society. This course builds on technical knowledge from previous foundations in Machine Learning and Neural Networks to provide a deep understanding of current AI platforms. Emphasizing hands-on experience in Generative Artificial Intelligence, students will learn to implement and train advanced AI models, including but not limited to transformers, diffusion models, and Large Language Models (LLMs). Additionally, the course will critically examine the ethical implications of AI, exploring the benefits, challenges, and potential risks associated with its deployment. Students enrolling in this course should have proficiency in Python programming, and a solid foundation in mathematics (including linear algebra and multivariable calculus) as well as statistics.
Instructor(s): E. Lo Terms Offered: Winter
Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 37200. Learning, Decisions, and Limits. 100 Units.
This is a graduate course on theory of machine learning. While ML theory has multiple branches in general, this course is designed to cover basics of online learning, along with basics of reinforcement learning. It aims to establish the foundation for students who are interested in conducting research related to online decision making, learning, and optimization. The course will introduce formal formulations for fundamental problems/models in this space, describe basic algorithmic ideas for solving these models, rigorously discuss performances of these algorithms as well as these problems' fundamental limits (e.g., minmax/lower bounds). En route, we will develop necessary toolkits for algorithm development and lower bound proofs.
Instructor(s): F. Koehler, H. Xu Terms Offered: Winter
Prerequisite(s): Requires linear algebra (at the level of CMSC 25300 or its equivalent), algorithms (CMSC 27200 or its equivalent) and probability (STATS 25100 or its equivalent). If not sure, consult with the instructor.
Equivalent Course(s): STAT 37201

DATA 37711. Foundations of Machine Learning and AI - Part I. 100 Units.
This course is an introduction to machine learning targeted at students who want a deep understanding of the subject. Topics include modern approaches to supervised learning, unsupervised learning, and the use of machine learning in estimating real-world effects. In principle, no previous exposure to machine learning is required. However, students are expected to have mathematical maturity at the level of an advanced undergraduate, including being comfortable with linear algebra, multivariate calculus, and (non-measure theoretic) statistics and probability. Assignments include programming in python (and pytorch).
Instructor(s): V. Veitch Terms Offered: Autumn
Prerequisite(s): Consent of Instructor unless graduate student in Data Science
Equivalent Course(s): CAAM 37711, STAT 37711

DATA 37712. Foundations of Machine Learning and AI - Part II. 100 Units.
Deep generative models have become a staple of modern machine learning research. This course is meant as an introduction to the way generative models are structured and trained: students will learn the mechanics of generative models as well as getting their hands dirty building them. We will discuss open questions for which we lack complete theoretical or empirical answers, with importance placed on analyzing, interpreting, and making arguments from necessarily incomplete empirical evidence. We will have a specific focus on Autoregressive Transformers and their use as Large Language Models (LLMs), but will also touch on Diffusion Models as well as Reinforcement Learning. The goal of this course is to get students to be proficient enough with the inner workings of deep generative models-along with the theoretical and empirical support for their design-to be able to understand and reason about cutting-edge research. This is an advanced machine learning course, and assumes a familiarity with basic machine learning concepts (generalization, overfitting, etc.) and techniques (regularization, stochastic gradient descent, etc).
Instructor(s): A. Holtzman Terms Offered: Winter
Prerequisite(s): DATA 37711; Consent of Instructor unless graduate student in Data Science
Equivalent Course(s): CMSC 37712

DATA 37784. Representation Learning in Machine Learning. 100 Units.
This course is a seminar on representation learning in machine learning. The core questions in this are: how do machine learning systems recover the structure present in real-world data, how can we expose this recovered structure to human analysts, and how does this help us in real-world applications? In this seminar, we will read and discuss papers from the modern research literature on these subjects. Students should have previous exposure to machine learning and deep learning.
Terms Offered: TBD
Equivalent Course(s): STAT 37784